Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 6 | Issue : 12 | Dec 2020

A COMPARATIVE STUDY ON SPAM EMAIL: DATA ANALYSIS BY VARIOUS CLASSIFICATION ALGORITHMS ALONG WITH JUSTIFICATION OF J48

Vutharkar Nagaveni[1] | Dr. Vimal Pandya[2]

[1]Computer Science, Rai University, Saroda, Gujarat, India.

[2]Computer Science, Director, Navgujarat College of Computer Applications,

ABSTRACT

Now a days email become one among the fastest and most economical and effective media of communication. Hence as increase of email users dramatically increase of spam emails during the past few years. The data mining classification algorithms are classified into categorize this email as spam or non-spam. During this paper, we conducted experiment within the WEKA environment by using three algorithms namely Naive Bayes, J48, Support Vector Machine (SVM) on the spam email dataset and later the three algorithms were compared in terms of classification accuracy. The in-depth analysis of the study and descriptions of the three classification algorithms is presented consistent with our data simulation results the J48 classifier outstanding performs than Naive Bayes and SVM in terms of classification accuracy level performance.

KEYWORDS: Classification, Accuracy, SVM, J48, Naive Bayes, WEKA.

I. INTRODUCTION:

Email is that the short kind of email correspondence and it's defined because the exchange of data through channel . Mostly emails come from different email addresses instead of being entered from the key board or electronic files stored on the disk or devices. Most mainframes, minicomputers, and therefore the emailing system are applied on the pc network. The term electronic message also can be written as Email or e-mail. Email address, which is required to send from and receive to email messages. the bulk of internet service providers provide a free email account to customers. Email has been tested to be one among the Internets preferred services; it's used for international communications. But, it's criticized for its insecurity, spam, also as viruses and malware being unfold through email attachments. E-mail offers how for web users to easily transfer information globally. E-mail presents an excellent way to send many commercials freed from charge for the sender, but the bad thing is that these days' emails are appreciably exploited. In general receiving the e-mail from unknown users comprises contents which are of not importance to the user. As a result, by these e -mails, many of users are getting cluttered with all unsolicited bulk e-mails also mentioned as “spam” or “unsolicited mails” (Vinod et al., 2013). Spam often causes unwanted information or bulk information to induce transmitted to email accounts. Spam mail might be a collection of electronic spam involving nearly identical messages sent to numerous recipients. Spam emails are conjoint and embrace malware as scripts or alternative executable file attachments to the browser net. Spam is waste of your time , space for storing and communication bandwidth. If spam continues to extend , it will be unmanageable within the near future to handle such huge spam.

II. RELATED WORK:

In this paper on Comparison of four email classification algorithms using WEKA used four sorts of classification algorithms for spam emails , which are namely, LAZY-IBK, Naïve Bayes, BayesNet, and J48, for his or her classification accuracy performance by using the WEKA environment. An experimental analysis, which was compares the four classification algorithms on the idea of parameters, like 'accuracy', 'precision', 'recall', 'F-measure' and 'false positive rate', to measure the performance of those four classification algorithms was performed and therefore the result was analyzed. This result reveals that J48 gave the most accurate results among the four algorithm.[1]

A study was conducted on four algorithms (J48, ID3, Alternating Decision Tree, and simple CART) for classification accuracy Spam datasets were run through the algorithms during a WEKA environment and it had been seen that J48 outperformed other algorithm three.[2]

In this author carried various classification algorithms of Hidden Naïve Bayes, Radial Basis Function (RBF) Network, Voted Perceptron, , Logit Boost, Rotation Forest, NNge, Bayesian Logistic Regression,Logistic Model Tree, REP Tree, Naïve Bayes, Multilayer Perceptron, Lazy Bayesian Rule,Random Tree and finally J48. This performance of were measured in terms of Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error, Receiver Operator Characteristics Area and Root Relative Squared Error by using WEKA tool.[3]

In this study e-mail data were classified as ham and spam email by using supervised learning algorithms with three different classifiers like Naïve Bayesian (NB) classifier, K-nearest neighbor (KNN) classifier and Support Vector Machine (SVM) classifier. The experiment was done by applying filtering on the classifiers and final result shows the difference between the classifier before and after applying filtering algorithm. The performance of the classification methods or algorithms are namely Naïve Bayes, SVM and KNN, true positive, false positive, precision, recall and F-measure were validated. There was a time difference is found in those classification algorithms. KNN and SMO algorithms are almost the most effective classifiers among the three before applying filtering algorithm. The Sequential Minimal Optimization (SMO) is employed to resolve quadratic programming (QP) problem that arises on the training of Support Vector Machines (SVM) , after applying filter. SMO algorithm is that the best classifier algorithm. [4]

III. MATERIALS (DATASET) AND PROCESSING METHOD:

In completing this research three steps were involved: Dataset Preparation, Pre-Processing and Application of varied machine learning classifiers and evaluating the performance of machine learning classifiers.

Dataset Preparation, Pre-Processing and Algorithm Application.

The Spambase dataset taken from the UCI Machine Learning Repository where the dataset has 57 attributes of various variable types in 4601 instances. The Spambase dataset is converted into .arff format (a format compatible for machine learning) supported by the WEKA tool for input file that was used for the analysis.

To classify the Spambase dataset, Naïve Bayes, J48, SVM were used and a 10 folds cross validation was used during this research. the selection of 10 folds was thanks to results obtained from broad tests on various datasets, with varying learning procedures, that have demonstrated that 10 is about the proper number of folds to induce the simplest gauge of error [8]. For cross-validation, a specified number of folds is chosen, the info is partitioned arbitrarily into 10 parts during which the class is represented in approximately an equivalent proportions as within the full dataset. Each partition is held to call and thus the training scheme trained on the remaining nine-tenths; then its error rate is processed on the hold-out set. Hence, the training procedure is carried out a total of 10 times on various training sets (each of which have tons in common). Finally, the averages of the 10 error estimates are taken to provide an overall error estimate.

IV. CLASSIFICATION METHODS APPLIED:

we usually consider the task of fraudulent e-mail detection as a classification task. Promising classification results will be achieved with the selection of representative features.In this section we will be discuss the classification algorithms used for detection of fraudulent or spam e-mails.

J48:

In the classification algorithms, decision tree method is one in all the famous methods because of its simplification and inductive nature. J48 technique is WEKA's implementation of C4.5 [5], a documented decision tree algorithm.J48 is an open source Java implementation of the C4.5 algorithm within the WEKA data mining tool. C4.5 is an algorithm accustomed generate a decision tree developed by Ross Quinlan. C4.5 could be a software extension

SVM:
Support Vector Machine (SVM) is widely used and regarded as state-of-the-art classification method for text classification.It has an benefit over others that it can work well on high dimensional feature set. SVM has another advantage that it can transform non-linearly separable data to a replacement linearly separable data by using kernel trick [6].Support Vector Machines (SVM) are supervised learning algorithms that are proven to perform betterthan another attendant learning algorithms. SVM may be a group of algorithms proposed by for solving classification and regression problems. SVM find an application providing solution to the quadratic programming problems which have inequality constraints and linear equality by differentiating the different groups by means of a hyperplane. It takes full advantage of the boundary.

Naive Baye's (NB):
NB [7] is another well know algorithm used for classification, which uses Baye's theorem. It calculates the probabilities of the feature values for every of the classification category and uses these probabilities to predict the class of the unknown instances.The Bayesian classification exemplifies a supervised learning technique and at an equivalent time a statistical technique for classification. It acts as a fundamental probabilistic model and allow us to seize ambiguity about the model in an ethical way by influencing the probabilities of the results. it's accustomed provide solution to analytical and predictive problems. Bayesian classification is called after thomas bayes (1702–1761), who proposed the algorithm. The classification offers practical learning algorithms and previous knowledge and experimental data will be merged. Bayesian Classification offers a beneficial viewpoint for comprehending and appraising several learning algorithms. It computes exact likelihoods for postulation and it's robust to noise in input data.

V. EXPERIMENTAL RESULT AND PERFORMANCE ANALYSIS:

The entire dataset was used for the experiment with 10 folds cross validation. The comparison of performance in terms of Accuracy, Precision, Recall, F-Measure etc.

This paper present a technique to classify mails supported three classifiers, i.e. J48, SVM, and Naïve Bayes. This classifiers were evaluated to separate spam from the e-mail dataset by using WEKA tool kit. The emails was identified as spam (1) or not spam (0), that reflected the attributes of the dataset of e-mail for spam filtering.

Table-1 A comparison of results between three algorithms

Classif cation algorithm	Accuracy (TP+ TN)/ (TP+ TN+ FP+ FN)	Recall TP/(TP+ FN)	Precision TP/ (TP+ FP)	FP RATE FP/(FP+ TN)	TP RATE (= RECALL)	F MEASURE 2*PR/(P+ R)	CLASS
NAÏVE BAYES	79.56	0.951	0.666	0.310	0.951	0.784	1
		0.690	0.956	0.049	0.690	0.801	0
J48	92.68	0.908	0.913	0.056	0.908	0.911	1
		0.944	0.940	0.092	0.944	0.942	0
SVM	90.48	0.831	0.918	0.048	0.831	0.873	1
		0.952	0.897	0.169	0.952	0.923	0

The analysis of the results demonstrated clearly that although J48 could be a very simple classifier which uses a decision tree, it gave the foremost accurate result to the experiment (92.68%).SVM also present good results with accuracy of 90.48% and better performance leads to other parameters too. But Naive Bayes is given accuracy of (79.56%), which is poor leads to comparison to other classfication.

In Ghada Hammad AL-Rawashdeh [1] study, the classification of mails completly supported four classifiers, i.e. BayesNet, J48, Lazy-IBK, and Naïve Bayesian. These classifiers were evaluated to separate spam from the e-mail dataset by using WEKA.The analysis of the results demonstrated clearly that albeit J48 may be a very simple classifier which uses a decision tree, it gave the foremost accurate result in his experiment as (85.06%).The LAZY-IBK also performs well with an accuracy of (84.003%) and BayesNet performs near by with accuray value of (83.3%).But Naive Bayes performs poor with an accuracy of (79.8%) rate.

In Aman Kumar Sharma [2] ,research is performed through the experiments, so as to work out the classification accuracy of four algorithms in terms of which algorithm better determine whether a specific email is spam or not with help of the data processing tool referred to as WEKA. Four algorithms namely ID3, J48, Simple CART and ADTree were compared on the idea of various percentage of correctly classified instances. of these four come under the classification methods of knowledge mining which makes a relationship between a dependent (output) variable and independent (input) variable by mapping the info points. In simple terms, classification problem refers to identifying an object as belonging to a given class as an example whether a selected mail is spam or non-spam.

It is clear form the simulation results that the very best classification accuracy performance is for the J48(92.76%) classifier for the spam email datasets containing 58 attributes with each 4601.

Furthermore Simple CART(92.63%) also showed similar results that were only slightly different from J48. ADTree (90.91%) and ID3 (89.11%) classifiers showed less accuracy as compared to the previous two mentioned. this means that J48 classification algorithm should be favored over Simple CART, ADTree and ID3 classifiers within the spam email application where classification accuracy performance is very important.

VI. CONCLUSIONS AND FUTURE RESEARCH:

This paper introduces a way to classify mails supported various classifiers, i.e. BayesNet, J48, SVM, Lazy-IBK, and Naïve Bayesian. These classifiers were evaluated to separate spam from the e-mail dataset by using WEKA.But analysis of the results demonstrated clearly that J48 is that the best performer for email classification.

However, an excellent deal of work is required within the future to verify the results for other algorithms. Future research includes improving an algorithm like Genetic algorithm, that there are many various mining and classification techniques. Also classified results might be utilized in Semantic Web by creating a modularized ontology supported classified result. additionally , different algorithms which aren't included in WEKA now also can be tested and experiments with various feature selection are often compared.

REFERENCES:

I. Ghada Hammad AL-Rawashdeh," Comparison of four email classification algorithms using WEKA",International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 2, February 2019.

II. Aman Kumar Sharma, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 5 May 2011.

III. Shafi'i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho, "Comparative Analysis of Classification Algorithms for Email Spam Detection ",I. J. Computer Network and Information Security, 2018, 1, 60-67 Published Online January 2018 in MECS (http://www.mecs-press.org/)DOI: 10.5815/ijcnis.2018.01.07.

IV. Elifenesh Yitagesu Desta* and Tekalign Tujo Gurmessa ," Analysis and result of classification algorithm on email classification",International Journal of Computer Engineering Research,SSN 2141-6494,Vol. 8(1), pp. 1-9, July-December 2019,

V. Quinlan JR. C4. 5: programs for machine learning. Morgan kaufmann; 1993.

VI. Joachims T. A statistical learning learning model of text classification for support vector machines. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM; 2001. p. 128–36.

VII. McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752; 1998. p. 41–8.

VIII. I. H. Witten and F. Eibe, Data mining : practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann Publishers, 2005.